

DOCUMENT RESUME

ED 118 606

TM 005 093

AUTHOR Pritchard, Robert D.; And Others
TITLE Development and Evaluation of an Objective Technique to Assess Effort in Training. Final Report.
INSTITUTION Institute for Organizational Behavior Research, Lafayette, Ind.
SPONS AGENCY Air Force Human Resources Lab., Lowry AFB, Colo. Technical Training Div.
REPORT NO AFHRL-TR-75-39
PUB DATE Oct 75
NOTE 51p.
EDRS PRICE MF-\$0.83 HC-\$3.50 Plus Postage
DESCRIPTORS *Ability; Comparative Analysis; *Incentive Systems; *Measurement; *Military Training; *Motivation; Objective Tests; Performance Tests; Predictive Ability (Testing); Technical Education; Test Validity
IDENTIFIERS Air Force; *Effort Measurement

ABSTRACT

This research explored the validation of a quantifiable, objective, and reliable method of measuring the amount of effort to be directly rewarded in incentive systems. A battery of relevant ability tests was given to a sample of Air Force trainees and to civilian subjects using a simulation of the course taught the Air Force trainees. Results showed that the simulation subjects were comparable to the Air Force subjects and that the ability test battery predicted performance equally well for both samples. The hard criterion of effort displayed wide variability, excellent reliability, and good construct validity. (Author)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

TM

AIR FORCE



HUMAN

RESOURCES

**DEVELOPMENT AND EVALUATION OF AN OBJECTIVE
TECHNIQUE TO ASSESS EFFORT IN TRAINING**

By

**Robert D. Pritchard
John H. Hollenback**

**Institute for Organizational Behavior Research
Lafayette, Indiana 47901**

Philip J. DeLeo

**TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230**

October 1975

Final Report for Period November 1973 - April 1975

Approved for public release; distribution unlimited.

LABORATORY

2

**AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235**

ED118606

**U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION**

**THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRE-
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY.**

TM005 093

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by the Institute for Organizational Behavior Research, Lafayette, Indiana 47901, under contract F41609-74-C-0010, project 1141, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Major Philip J. DeLeo, Instructional Technology Branch, was the contract monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

Approved for publication.

HAROLD E. FISCHER, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TR-75-39	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DEVELOPMENT AND EVALUATION OF AN OBJECTIVE TECHNIQUE TO ASSESS EFFORT IN TRAINING		5. TYPE OF REPORT & PERIOD COVERED Final November 1973 - April 1975
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert D. Pritchard John H. Hollenback Philip J. DeLeo		8. CONTRACT OR GRANT NUMBER(s) F41609-74-C-0010
9. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Organizational Behavior Research Lafayette, Indiana 47901		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 11410103
11. CONTROLLING OFFICE NAME AND ADDRESS Hq Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE October 1975
		13. NUMBER OF PAGES 50
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Technical Training Division Air Force Human Resources Laboratory Lowry Air Force Base, Colorado 80230		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) technical training incentive systems student performance effort measurement		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This research explored the validation of a quantifiable, objective, and reliable method of measuring the amount of effort to be directly rewarded in incentive systems. A battery of relevant ability tests was given to a sample of Air Force trainees and to civilian subjects using a simulation of the course taught the Air Force trainees. Results showed that the simulation subjects were comparable to the Air Force subjects and that the ability test battery predicted performance equally well for both samples. The hard criterion of effort displayed wide variability, excellent reliability, and good construct validity.		

SUMMARY

Problem

The problem addressed by this research concerned how to award incentives for achievement in training. A difficulty with most incentive systems is that they provide the greatest payoff to high ability students. In fact, in traditional incentive systems, lower ability students may not get rewarded at all, regardless of how hard they try. The present study was part of a program to tailor incentives to the capabilities of each individual student. The major purpose, therefore, was to develop and evaluate a method of objectively measuring the effort exerted by a student in a technical training context. One particularly desirable characteristic of such a measure would be its potential usefulness in an effort-based incentive system.

Approach

The literature concerning physiological, rating, behavioral, and computational techniques for measuring effort was reviewed. As demonstrated by this review, many of the traditional measures have serious limitations. The approach taken in this study was based on the assumption that performance is largely a function of ability and motivation (effort). A logical consequence of this assumption is that a measure of effort can be obtained by partialling out the influence of ability on performance. Thus, a residual score derived in this manner would reflect what level of effort an individual was exerting.

For the purpose of evaluating this derived measure of effort, an 8 1/2 hour section of an Air Force technical training course was selected for study. This section was taken from the Aircraft Electrical Repairman Course (3ABR42330), conducted at Chanute AFB, Illinois. Following an analysis of the course material, a battery of relevant ability tests was given to a sample of Air Force trainees in the target course. Regression equations predicting performance on the course material were then developed and cross validated.

Using civilian subjects whose personal characteristics approximated the Air Force population, a simulation of the selected course was developed. These subjects took the same ability tests and worked on the same materials as did the Air Force subjects. To assure variability in effort, three pay systems were used - hourly, piece-rate, and variable ratio/variable amount. A second set of regression equations predicting performance was developed and cross validated for the simulation sample.

Derived effort scores were then calculated for the simulation subjects using both the Air Force generated weights and self-generated weights by subtracting from actual performance the performance level predicted on the basis of ability. Finally, the derived effort scores of the simulation subjects were correlated with a hard criterion of effort based on a photographic record.

Results

Results of the study showed that the ability test battery predicted performance equally well for both samples. The hard criterion of effort displayed wide variability, excellent reliability, and good construct

validity. The derived effort measure showed moderate correlations with actual effort. When the analyses were done with high and low ability subjects separately, the correlations were larger.

Conclusions

It was concluded that the derived effort index would not be adequate as an index of a single individual's effort, but could be quite useful in assessing differences in effort between groups. A number of specific practical applications were discussed. Calculation of derived effort scores was recommended for (1) the award of incentives to groups, (2) the award of incentives to lower ability students (3) comparing the motivational characteristics of different courses or blocks within courses, (4) feedback to students and instructors about group effort, and (5) goal setting.

PREFACE

This research was conducted under Project/Task/Work Unit 11410103, Psychological Factors in Instructional Systems Design.

We would like to thank the members of the research staff who made many valuable contributions to the project: James R. Terborg, Blair Clark, Lee Stepina, Gail Schmaltz and Francine Lamattina.

We would also like to thank the Air Force personnel at Chanute Air Force Base for their extensive cooperation: Mr. R. Mitchell, Mr. H. Mull, Lt. D. Gillman, and Mr. G. Scharf.

Table of Contents

Introduction.....	5
Review of the Literature.....	7
Conclusions from the Literature Review.....	11
Methods and Procedures.....	13
Overview.....	13
Selection of Task Material.....	13
Ability Testing.....	14
Dependent Variables.....	15
The Work Simulation.....	15
Results.....	18
Overview.....	18
Comparison of AF and Simulation Subjects.....	18
Generation of Regression Equations.....	19
Generation of Derived Effort Scores.....	23
Evaluating the Hard Criterion of Effort.....	24
Predictive Validity of the Derived Effort Measure.....	26
Summary of the Results.....	30
Discussion and Conclusions.....	31
References.....	36
Appendices.....	39
A. Advertisement used in recruiting subjects.....	39
B. Pre-employment electricity test.....	41
C. Sample of an appraisal.....	43
D. Effort Questionnaire.....	46

Introduction

The goal of training in the Air Force is to maximize each individual's contribution to the Air Force mission. To accomplish this goal, Air Training Command (ATC) attempts to train each man to accomplish specified criterion objectives consonant with field requirements. The issue thus becomes one of optimizing performance in the training setting and ultimately in the field. Performance can be thought of as being composed of two major components, ability and motivation (Vroom, 1964). Clearly, other classes of variables influence performance, but most would agree that ability and motivation are extremely important components of performance (Campbell, Dunnington, Lawler & Weick, 1970). This argument implies that to maximize performance, one could maximize ability and maximize motivation. The ability component can be dealt with by giving remedial instruction to low ability students (e.g., remedial reading courses) and by selecting instructional strategies to fit each individual's abilities and traits. This leaves the issue of motivation. One of the approaches to the motivation issue has been to investigate the feasibility of using incentive motivation techniques in an ATC training environment. (e.g., Pritchard, Von Bergen and DeLeo, 1974).

The classical approach to incentive motivation has been to give valued rewards contingent on actual performance in some task. The important point is that rewards are given on the basis of performance. Thus, this classical approach would suggest that incentives be offered to airmen on the basis of their scores on exams and/or their speed of finishing sections of the course. However, there is a problem with this approach. Theories of human task motivation which deal with incentive motivation (e.g., Vroom, 1964; Porter and Lawler, 1968; Lawler, 1971) talk about three components that influence motivation: (1) Valence of rewards: the value the individual places on incentives, (2) Performance - reward instrumentality: the perceived degree of connection between performance and obtaining the rewards, and (3) Effort - performance expectancy: the perceived degree of relationship between a person's level of effort and his level of performance.

Only the first two of these components are generally considered by a classical incentive motivation approach. That is, valued rewards are identified and made contingent on performance (e.g., score on exams and speed of finishing.) This ignores the issue of effort - performance expectancy. Specifically, high performance may be seen as highly valuable, but if a person sees no relationship between his level of effort and attaining high performance, he will not be motivated to attain high performance. For example, if a man were offered \$10,000 to pick up a 2000 pound block of concrete, he would not attempt it even though the value of performing the feat would be very high. He would perceive that no matter how much effort be expended he could not pick up the weight. Taking a more realistic example, an airman in tech school might value a 3-day pass very highly but feels that being a top performer would be impossible for him. Thus, the pass would not motivate him. The problem is even greater when we consider that it is the low ability student who sees little chance of being a top performer, yet it is precisely this student we wish to motivate.

Thus, what is additionally necessary is to deal somehow with the relationship between effort and performance. This could be accomplished, at least in principle, in a rather simple way. If one were to maximize the contingencies between effort and rewards, students of any ability level could be motivated. That is, if an incentive system were so structured as to give valued rewards for high levels of effort, all students should be motivated.

The purpose of the research described here was to develop and evaluate a technique for objectively assessing effort. If this could be done in an economical and objective way, it would become quite easy to give incentives on the basis of effort.

Aside from the general issue of giving rewards on the basis of effort, this research ties in directly with the development of an Advanced Instructional System (AIS) by the Air Force Human Resources Laboratory. The AIS is a computer-managed systems approach to training. One of the central requirements of the AIS is that training packages be individualized. This individualization covers not only the type of training an individual receives but also attempts to maximize the actual motivation of each student. Developing a system which assesses student effort, and which could be used to award incentives is clearly consistent with the philosophy of individualizing the training package.

Review of the Literature

Definitions and Measurements of Effort¹

Techniques for measuring effort are as numerous as definitions of the term. For example, effort has been defined as a determinant of motivation (Atkinson, 1964); one component of motivation (Yacorzynski, 1942); the equivalent of motivation (Farquhar, 1963); and a measure of motivation (Lawler and Porter, 1967). Each of these definitions may indeed be appropriate depending upon the context in which the term is used. Regardless of the context, we would argue that effort is closely associated with the construct of motivation. As such, effort becomes a critical construct in the study of motivated behavior.

In this section, various techniques for measuring effort will be described and evaluated. The techniques to be discussed fall into four categories: physiological, rating, behavioral, and computational.

Physiological Techniques

A very fundamental approach to the measurement of effort involves the use of physiological indicators. From this perspective, motivation or effort, is viewed as a state of general arousal (Leukel, 1968). Generally, the level or degree of effort is determined by the activity of the central or autonomic nervous systems. McClelland (1955) lists several indices which might reflect this activity level:

1. Energy expenditure (basal metabolic rate)
2. Autonomic activity (skin conductance)
3. Thresholds (reaction time)
4. Muscle activity (eye movements and action potentials)
5. Central excitation level (determination of the frequency in cycles per second at which a flickering light fuses)

Other researchers have proposed additional physiological indices. Circulatory changes such as pulse rate (Bitterman, 1945) and pulse pressure (Lovekin, 1930) have been used to indicate effort expenditure. Conflicting reports suggest, however, that there is not an established technique for estimating effort by these changes in the circulatory system.

Measurements of muscular tension have also been used as indices of effort (Davis, 1939; Ryan, 1947; Solomon, 1946). Further, Scott (1960) proposed auditory flutter fusion (the frequency in cycles per second at which a fluttering sound fuses) as an index of effort.

A major problem with physiological indices of effort seems to be that such measures are confounded by a variety of factors - time of day that the measure was taken, amount of sleep that subject had, amount of food consumed, task difficulty, fatigue, etc.

¹This review is based heavily on two other reviews. (Ramby, 1973 and Mayo, 1974.) The majority of the two reviews were done under support of previous Air Force Human Laboratory contracts; R.D. Pritchard, Principal Investigator.

Beyond the substantial problem of confounding variables, there exists the difficult, obtrusive, and expensive nature of obtaining physiological measures. It would appear that frequent readings using rather obvious and expensive equipment are necessary to obtain good measures. Even when such indices are obtainable, further research is needed to determine the accuracy and applicability of such techniques in a field setting.

Rating Techniques

Observations by trained experts such as clinicians or time-study men have been used to assess effort. Early judgmental indices were used to rate a person's work pace as being faster, or slower than "normal" (Strauss and Sayles, 1960). For example, time studies (Barnes, 1940) require the observer to time each separate operation of the task and rate the worker on skill, effort, consistency, and the conditions under which the study was conducted. These ratings are then numerically adjusted to correct for differences in observation time. Presgrave (1945) developed a time-study technique wherein measures of effort were determined by variations in the speed to complete a task. It would seem that such a method based solely on speed would also include the effects of skill, and therefore not present a clear assessment of effort. By classifying workers according to skill level, Ryan (1947) refined the criteria for determining levels of effort from observers' ratings.

One difficulty with observer ratings of effort is that raters may be responding to performance rather than effort. Braunstein, Braunstein, and Blumfield (1965) assessed the relationship between an overall observed rating of effort and various measures of actual performance. The authors found that effort ratings were only related to three of the six performance measures. They concluded that the raters may indeed have been responding to something other than level of performance. To insulate effort ratings from the effects of knowledge of performance, some researchers (Mitchell, 1966; Hackman & Porter, 1968; and Schneider & Olson, 1970) have devised instruments requiring the raters to focus on effort rather than performance. Such attempts to separate performance from effort seem to be an improvement over previous scales.

Since judgments of effort certainly have a subjective component, self-ratings have also been used to measure effort. Obviously, judges or raters are not required, nor is any mechanical device needed to make this type of effort determination. Typically, self-ratings simply require an individual to rate himself on some type of scale according to the amount of effort put into a specific task or job.

Thorndike (1913) asked subjects to rate their effort on various parts of a learning task. Furst (1966) developed an effort scale and a measure of motivation. Subjects rated themselves on a five-point scale for nine effort statements. Furst found that this effort rating instrument correlated higher with a measure of motivation than did an achievement measure. Employee attitude scales pertaining to different aspects of the job have also been used to measure effort (Lawler and Porter, 1967).

Because of their subjective nature, both observer and self ratings of effort may be contaminated by the same factors that bias all rating scales: for example, response set, halo effect, and leniency effect. This problem, in addition to questionable validity and reliability, suggests that a more objective technique for measuring effort would be desirable.

Behavioral Techniques

Researchers in learning theory have used certain behaviors of human and infra-human organisms as indices of effort. One approach establishes a motivational state by exposing the organism to a set of antecedent conditions, and then observing the change in the organism's behavior. This behavior change is considered a measure of motivation. For example, in the case of hunger, motivation would be operationalized by subjecting laboratory animals to various degrees of food deprivation. The rate of response behavior (bar pressing) would be the measure of motivation or effort. Obviously, such a rate of response measure is of little value when dealing with the complex nature of human effort.

Another approach, more useful with humans, is simply to measure behaviors that appear to correlate with effort. Davis (1939) found certain movements of the right arm to be associated with the effort experience in the solution of arithmetic problems. Luchins and Luchins (1954) employed a mirror-tracing task and observed that fidgeting and sweating accompanied high levels of mental effort. Yacorzynski (1942) found some evidence to indicate that time taken to complete a given task was related to effort. Such behavioral indices are subject to a variety of problems. Measures like arm movements, fidgeting, and sweating are potentially confounded by any individual or situational variable that also might cause such behaviors. Task completion time as an effort measure does not account for ability differences. Thus, it seems that behavioral measures are subject to several specific problems, as well as the general problems of rater bias and unreliability. Such problems significantly reduce the desirability of using behavioral indices of effort.

Computational Techniques

Another general approach to the measurement of effort is based simply on computationally deriving an effort score. Educators interested in obtaining such effort scores typically use the ratio or difference between achievement and intelligence scores. The simplest of these indices of effort is the Accomplishment Quotient (Pintner, 1920.) The AQ is the ratio of an individual's actual rate of educational progress (Educational Quotient) to the potential rate of progress (Intelligence Quotient). According to Haven (1931), this measure of effort evaluates the accomplishment of an individual in terms of his own ability. Deviations above and below 1.00 indicate the degree of effort expenditure and ability utilization in various performance tasks.

The Efficiency Ratio (ER) has also been used as an index of effort (Ford, 1931). Statistically it is :

$$ER = S/Av(IQ/100) \times 100,$$

where S is the individual's score on an achievement test and Av is the Average of an experimental group on the same achievement test. A resulting score of 100 indicates that the individual is exerting normal or average effort. Scores above 100 reflect above average effort, whereas scores below indicate below average effort.

The Effort Quotient (FQ) also used a ratio techniques to measure effort (Tsao, 1943).

$$FQ = (E/\text{predicted } E) \times 100,$$

where E is an individual's educational score, and predicted E is the predicted value of the educational score based on an intelligence measure. More specifically, the regression equation used to predict an individual's educational score is:

$$\text{pred } E = M_e - b_{ei}M_i + b_{ei}I = b_{ei}I + K,$$

where M_i and M_e are the mean values of intelligence and achievement, respectively; and I is the known individual intelligence score. The final component b_{ei} is determined by:

$$r_{ei}S_e/S_i$$

where r_{ei} is the correlation between the intelligence and achievement scores; and S_i and S_e are the standard deviations of the intelligence and achievement scores, respectively. If the resulting value for the FQ is 100 (actual educational score equals predicted score), then according to Tsao, the individual is exerting normal effort. Values of FQ above or below 100 reflect higher or lower amounts of effort, respectively.

A common difference technique used to assess effort is the Effort Score (McCall, 1930).

$$F = T_e - T_i + 50,$$

where T_e and T_i are T scores on an achievement and intelligent test respectively (a T score is 1/10 of the standard deviation of the experimental group on the particular measure). An individual whose T_e score is equivalent to his T_i score is said to be exerting normal effort. Scores above and below 50 reflect various degrees of effort.

In a review of the previously mentioned techniques for measuring effort, Tsao pointed out that the AQ score, the F score, and the ER score fail to consider the correlation between intelligence and education. He concluded, therefore, that each of these techniques gives a biased estimate of effort.

Similar techniques have been used in other areas of Education. One other common approach to determining over- and underachievement, is the derivation of a discrepancy score. This particular technique is computed simply by subtracting the aptitude score for an individual from his achievement score. The resulting residual is a measure

of over or under-achievement. That is, a negative discrepancy score represents under-achievement, and a positive one over-achievement. Because this crude difference score leads to a systematic negative bias for those individuals high in aptitude, and whereas a positive bias will result for those low in aptitude, Thorndike (1963), suggests that achievement be predicted from aptitude on the basis of the known correlation between the aptitude measure and the achievement measure (this is similar to the FQ technique). This prediction or regression equation will give the average achievement score for individuals at any given aptitude level. The predicted value will then be an unbiased estimate of achievement.

Summary of Effort Measures

The techniques discussed, particularly rating scales, tend to be too subjective to adequately assess effort. Of concern are the effects of rating biases which may serve to contaminate the measure. That is, raters may be responding to a dimension other than effort. Further, the question remains relatively unanswered as to whether physiological measures or rate of responding measures do indeed reflect effort. These techniques focus only on output, and therefore may not be completely representative of an individual's effort.

Given that a motivated individual is one who exerts both physical and mental effort, effort appears to be defined along more than a single dimension. What is needed then, is an approach which is reflective of both the physical and mental aspects of effort, and at the same time is an objective and valid representation of the construct. The work of Tsao with the Effort Quotient, and the work of Thorndike with over- and under-achievement appear to come closest to satisfying these criteria. Although Thorndike's measure is not conceptualized in terms of effort, it does afford an indirect and less subjective approach to the measurement of over- and under-achievement. Further, this measure has been connected to motivation (effort) by a few authors (Appelzweig, Moeller and Burdick, 1956; Mayo and Manning, 1961; and Farquhar, 1963). In each of these studies, the terms over- and under-achievement have been used as the operational definitions of high and low motivation. That is, this discrepancy or residual score was considered a measure of motivation.

The Effort Quotient (FQ), while similar Thorndike's technique, uses a ratio approach for assessing effort, rather than a difference score. Further, since FQ score was developed in terms of effort it may also be valuable in the construction of an effort measure.

Conclusions from the Literature Review

Clearly there have been many approaches to measuring effort. Most of them, however, have serious limitations for use in an Air Force training context. Ratings and physiological measures would not only be difficult to collect, but their validity is questionable.

The approach taken by the present research is to deal with the derived effort measure originally developed by Pritchard, Von Bergen, and Deleo, 1974. The approach assumes that performance is largely a function of ability and motivation (effort). Given this assumption, which can be empirically tested, one can get a measure of effort by partialling out the influence of ability on performance. The residual could then be considered a measure of effort. Specifically, one could generate a prediction equation from ability test data which taps relevant ability domains of a given training course. Once this equation is generated, it could be used to calculate a predicted score (a level of performance and/or speed of completion) of a given individual on a given segment of a course. Conceptually, this predicted score would be the mean actual performance of a group of students with similar sets of ability scores. Thus, since ability is constant, variations in performance should be due to variations in effort.

This predicted score for each individual can then serve as the mechanism for determining effort. This would be accomplished by subtracting the predicted score from the student's actual score. This residual could then be considered as a measure of effort, and the higher the score, the more effort the student exerted.

It is obvious that the entire system rests on the assumption that technical school performance is largely a function of ability and effort, and that when ability is partialled out the remaining variance in performance is highly saturated with variance in effort. It is precisely this assumption that the present research was designed to test. Before discussing the research plan, however, let us consider some of the advantages of an incentive system which gives rewards based on effort.

First, such a system is individualized. Each person gets a predicted score based on his own pattern of abilities. Second, the system would reward effort rather than performance. The advantages of this approach were discussed earlier. Third, the system equalizes the incentive system in that all students have an equal chance to earn incentives. Related to this is the advantage that all students should be motivated to high effort, not just the high ability students. Fourth, the measure of effort could serve as a very useful overall measure of the effectiveness of changes in the course. Specifically, innovations and constant changes will undoubtedly be made in a course. The problem is how to assess the impact of these changes. Our own experience has shown that the ability levels of students entering any given technical school course can vary greatly over a short period of time. This makes assessing the effects of any changes very difficult without partialling out ability somehow. The system proposed here does this automatically and thus is directly interpretable. Other less obvious advantages would include the identification of problem students, and use for counseling purposes. For example, the system would easily identify a student of high ability who was barely passing the course (low effort). Finally, the system would be useful for other motivational applications. For example, the target score could be used as the basis for various types of goal setting procedures.

Method and Procedures

Overview

Initially, it was necessary to select an ongoing AF technical training course to use as the target course and examine the course materials in order to assess the abilities required by that material. Based on this analysis, a battery of ability tests was formed to tap relevant abilities. These tests were then given to a sample of AF trainees in the selected course. Regression equations were then developed to predict performance in the target course. The next step was to develop a simulation of the selected course using civilian subjects. These simulation subjects took the same battery of ability tests, and worked on the same training materials as did the actual AF subjects. A second set of regression equations was developed for the simulation sample.

Derived effort scores were then calculated using both the Air Force generated weights and the self generated weights by subtracting from actual performance the performance level predicted on the basis of ability. The final step consisted of comparing the derived effort scores of the simulation subjects to a hard criterion of effort which was based on a photographic record. The greater the relationship between the derived effort measure and the hard criterion of effort, the better the derived effort could be said to be measuring effort.

Selection of Task Material

Four criteria were used in the selection of task material. The first was that it be part of an ongoing Air Force technical course. This was necessary to be able to maximally generalize to the Air Force context. Second, it had to be self-paced. Since the AIS, for example, is a self-paced system, the findings would be more usable if a self-paced course was utilized in the simulation. Third, it had to utilize programmed texts. It was not feasible to have a fully trained instructor in the simulation, and thus the use of programmed texts was felt necessary. In addition, the AIS will use programmed materials. Finally, for logistical reasons, it had to be a section of materials which did not require extensive training equipment.

A number of possibilities were examined, and ultimately the Aircraft Electrical Repairman (AER) course (3ABR42330) was selected. This is a self-paced course utilizing programmed texts which covers the fundamentals of electricity and the maintenance and repairs of electrical systems in aircraft. Although the complete course requires extensive training equipment and skilled instructors, the first section of the course covers more basic material and does not require equipment.

The section of the AER course used in this study consisted of the following programmed texts:

- Aircraft Familiarization
- Elements of Physics and Mechanics
- Electron Theory

Magnetism

DC Generation and Basic Circuit Symbols and Terms

Series Circuits Wiring Diagrams

Because this 8½ hour section occurred early in the sequence of course work, students were not required to use any training equipment. As a whole, the programmed texts were introductory in nature and were designed to provide basic knowledge and background necessary for the completion of subsequent course work.

The specific learning objectives provide an excellent description of the content area of each programmed text. In Aircraft Familiarization, students were required to identify aircraft components, aircraft movements, and the direction of aerodynamic forces from diagrams. Also required was a working knowledge of the alphanumeric aircraft designation system.

Elements of Physics and Mechanics was used to train students in the principles and methods of using simple machines. Also required was a knowledge of the causes and controls for various types of corrosion. Students had to identify the effects of pressure and temperature changes on solids, liquids, and gases.

Successful completion of Electron Theory required an understanding of subatomic particles as well as the principles, symbols, and measures associated with voltage, resistance, current, and conductances.

The programmed text, Magnetism, dealt with characteristics of permanent magnets and electromagnets. Students were required to grasp basic concepts of magnetism and had to identify electromagnetic effects from diagrams.

DC Generation and Basic Circuit Symbols and Terms was used to teach students the basic operation of electric components associated with direct current generation. Additionally, the text was used to train students to recognize certain electrical symbols and terms.

The final text, Series Circuits Wiring Diagrams, was used to train students in the use of electrical station numbering systems. Terms and procedures associated with this system were stressed.

Ability Testing

Once the materials were selected, they were carefully examined to identify the ability dimensions required by the materials. This was done on an intuitive basis. Ultimately, a battery of five tests were selected.

These tests were selected because they were: 1) designed for standardized group administration, 2) highly reliable and valid, and 3) relevant to the ability requirements of the task material. The Otis-Lennon Mental Ability Test was used to measure general intelligence. General abilities related to logic, mathematics, and vocabulary are measured via this instrument. Such abilities are considered relevant to nearly any learning task.

The Paragraph Meaning Test was selected from the most current version of the Stanford Achievement Test Battery. This test was selected to tap

reading comprehension skills. The format of this test closely parallels the format of a programmed text - a paragraph is presented and related questions and completion items follow immediately. It was reasoned that the ability to comprehend and respond to a small segment of written material would be very relevant to a programmed learning task.

While progressing through the task material, the subjects were also required to interpret figures and diagrams. The Study Skills of the 1952 version of the Stanford Achievement Test Battery was selected since it dealt primarily with the extraction, synthesis, and interpretation of information presented graphically and diagrammatically.

Also important to completing the programmed texts was very basic mechanics and an understanding of topics such as simple machines, magnetic lines of force, and electron flow. To measure related abilities, the Mechanical Reasoning and Space Relations sections of the Differential Aptitude Tests were administered.

Once selected, the battery was administered to a sample of AF trainees in the AER course. This was accomplished by giving the three-hour test battery to students as they entered the course. A member of the research staff administered the battery to students on their first day of class. Trainees were told that the tests were important in that they tapped abilities relevant to doing well in the course, and that they should do their best on the tests.

Dependent Variables

Final data collection in the Air Force context consisted of performance data over the first six PTs for those trainees who had taken the ability test battery. Data were collected on three variables. The major dependent variable was time-to-complete the first six PTs. Also, data on scores (percent correct) for each PT were collected. Finally, when the trainee had finished the six PTs he was given a specially designed comprehensive test over all the material covered in the six PTs.

Data on the first two variables were collected by having the instructor record, on a specially designed form, the time a student started a given PT, when he took the appraisal test for that PT, his score on the appraisal, and the time he started the next PT. If the student failed to pass the PT, this was also recorded. Thus, it was possible to calculate total time spent on the PTs, as well as mean appraisal score.

The comprehensive test was developed especially for this project to cover all the material in the six PTs. Sixty items were initially compiled. Many items were newly developed while others paralleled items found in course appraisals and criterion tests. These items were evaluated for accuracy of content by AER course instructors and then administered to 32 students in the program. Based on an item analysis of their responses, a final 55-item version of the comprehensive test was developed.

The Work Simulation

The second phase of the research was to design a simulation of the AER

course, collect the same ability and performance data, and obtain a hard criterion of effort.

Task Material

The first six PTs of task material used in the simulations were identical to those used in the actual AER course. Copies of the actual PTs were made, and the same exact tests were used. Since the major dependent variable was the time it would take to finish the six PTs it was important that every subject actually finish the material during the simulation. Thus, it was arranged so that, based on available Air Force experiences, everyone should be able to finish in the 20-hour working time scheduled for the simulation. However, it was anticipated that most subjects would finish in less than 20 hours, and the better students were expected to finish much more quickly. Consequently, additional task material was generated for use when the subject had finished the first six PTs. This material was based on published programmed texts in electricity and electronics (New York Institute of Technology 1963, 1964) and had been used successfully in a similar setting by Pritchard, Leonard, Von Bergen, and Kirk (1974). However, for the purposes of this report, only the data from the first six PTs are relevant.

Subjects

It was felt important that the subjects selected for the simulation be as similar as possible to the trainees in the AER course. Thus, an attempt was made to recruit subjects of the same age range and general ability level as those in the AER course.

The simulation was actually composed of three separate data collections, in three separate cities in Indiana. Approximately two weeks before the simulations was due to start, advertisements were placed in local newspapers and flyers were distributed in the area describing the job and telling subjects where to report. (See Appendix A for a copy of the advertisement.) It was planned to have 20 subjects in each of the three data collections, and the advertisements were quite successful in that each condition, more than twenty people showed up for the job.

Procedure

Once the subjects arrived, they were given an application blank and the job was described to them. They were told that we were interested in a new method of technical training which involved programmed texts, and that they would be working with these programmed texts in electricity and electronics. It was pointed out that no special skills were required, and that it was not necessary that they have any previous experiences or knowledge of electricity or electronics. At that point, anyone not interested in pursuing the job was told that they could leave. No one, in fact, actually left.

They were then told that, as the ad stated, we only needed 20 people. Since more than 20 were present, some of them could not be

hired. They were then given a short test of knowledge about electricity and electronics. This test dealt with questions which would not generally be familiar to someone without some background in this area. Examples were "Define static discharges, magnetic permeability etc." (a copy of the test appears in Appendix B.)

Although subjects were advised to do as well as possible on the test, its actual purpose was to eliminate those people who had some knowledge of electricity and electronics.

Anyone who got more than one answer correct on this test was eliminated. Of those applicants remaining, 10 males and 10 females were randomly selected. The rest of the applicants were thanked and dismissed.

The remainder of the first day consisted of giving the subjects the battery of ability tests used in the Air Force sample, explaining the task in detail, and giving subjects some practice working on a sample PT.

In order to assess the utility of the derived effort measure it was necessary to assure that some variability in effort was, in fact, present. To this end, three conditions were utilized, varying in the type of pay system employed. The first was an hourly system whereby subjects received \$2.00 per hour. The second was essentially a piece rate. Each PT was given a dollar value, and when the subject passed the appraisal for the PT, he received that amount of money. The more PTs finished, the more money he would earn (All money earned was paid at the end of the week). This "value" of each PT was based on the data from the hourly condition. If, on the average, hourly subjects took e.g. 2 hours to finish a given PT, this value was multiplied by the \$2.00/hour rate to get the "value" of that PT. In this example, it would be \$4.00. Thus, if subjects in the piece rate condition worked at an average pace, they would make \$2.00 per hour. If they worked faster, they would earn more.

The third condition was similar to the piece rate or fixed ratio (FR) condition in that pay was contingent on performance, but the actual pay schedule was different. The third condition utilized a variable ratio-variable amount (VRVA) schedule. In this condition, subjects did not know how much a given PT was worth since its "payoff" varied from \$0 to \$6.00 times the number of hours taken to complete it for the hourly condition. Thus, for a PT which took 2 hours to complete in the hourly sample, subjects in the simulation sample could earn \$0 through \$12.00. The determination of which level of pay they actually received was random, but set so as to average \$2.00 per hour if performance equalled that in the hourly sample.

The three conditions were run independently in three different cities and subjects in one condition were exposed only to that condition. The system was explained the first day, and subjects worked the following four days under the appropriate system. Including the first day used for testing and orientation, subjects were in the simulation for five days, five hours per day.

At the start of the second day, all subjects were given the first PT and told to start. When they felt they knew the material, they approached the instructor and were given the appraisal for the first PT. (A sample appraisal appears in Appendix C.) When completed, it was scored by the instructor. If the subject passed the appraisal, he was given the second PT. If he failed the test (75% correct was the criterion for passing) he was told to re-study the PT. When he felt he was ready he re-took the

first appraisal. Actually, this was another form of the appraisal. He continued this procedure until the appraisal was passed. When a subject finished the set of six PTs, he was given the comprehensive test covering the whole set of PTs, the same test given to the Air Force trainees. The subject was not required to reach any set criterion score on this test. Subjects were then given a brief interview and completed a questionnaire (See Appendix D.) Once they had completed this they started on the "new" material and worked on it for the rest of the week. Throughout the entire work week subjects could take breaks whenever they wished, for as long as they wished. A separate break area was provided, and coffee, soft drinks, and doughnuts were available.

The hard criterion of effort was based on a photographic record. Each working day two 8 mm movie cameras took single frame pictures of the working area. Each subject's work place was clearly visible. The cameras took one frame every six seconds for the entire day. The cameras were clearly visible to the subjects, and actually made an audible click when each frame was taken. However, subjects quickly adapted to the cameras and by the second day, when the effort data were actually collected, there was absolutely no evidence that the subjects were paying any attention to the cameras. As will be discussed below, the measure of effort consisted of the percent of time the subjects actually spent working on the task material.

Results

Overview

Analysis of the data consisted of four basic steps: 1) generating the regression equations, 2) computing the derived effort scores, 3) evaluating the hard criterion of effort, and 4) examining the predictive validity of the derived effort scores. Before turning to these topics, we shall first address the issue of the comparability of the Air Force, and simulation subjects.

Comparison of Air Force Simulation Subjects

Table 1 presents comparisons of the two subject groups in terms of age, IQ and the five ability tests. These data indicate that the Air Force trainees tended to be slightly older than the simulation subjects. This is due primarily to the fact that the Air Force group included some trainees who had been in the service for some time, but had returned for retraining in this career field. In fact, 79% of the Air Force trainees were between 17 and 20 years old. Thus except for the retrainees, the groups were comparable in age. Although no actual education data are available our experiences with this course (Pritchard, DeLeo and Von Bergen, 1974) suggests that almost all of the Air Force trainees had some high school and about 70% had completed high school. This corresponds closely to the amount of education of the simulation subjects.

The Table also indicates that in terms of IQ and Paragraph Meaning, the ability level of the simulation subjects was higher. However, the two groups were about equal in terms of Mechanical, Spatial and Study Skills ability. For both groups, substantial variability existed in ability.

Table 1
Comparison of Air Force and Simulation Subjects

	AF Subjects (N=187)			Simulation Subjects (N=60)		
Variable	\bar{X}	S.D.	Range	\bar{X}	S.D.	Range

Variable	\bar{X}	S.D.	Range	\bar{X}	S.D.	Range
Age	19.3	2.4	17-32	17.4	.6	17-19
IQ	92.6	9.2	70-127	107.9	15.9	63-137
Paragraph Meaning	80.6	21.9	0-124	101.5	21.2	48-127
Mechanical	24.6	4.6	9-35	24.8	4.9	14-36
Spatial	15.4	5.6	0-28	16.5	6.5	5-28
Study Skills	23.2	5.2	9-34	26.1	4.9	11-33

Generation of the Regression Equations

The basic strategy here was to attempt to predict performance on the training materials from the ability data. This predicted performance score would then be compared to the actual performance score to obtain the derived effort measure. The optimal procedure would be to generate a regression equation on the Air Force trainees and apply this equation to the simulation subjects. This would eliminate the possibility of capitalizing on chance that would exist if the equation was based on data from only the simulation sample. Thus, the primary regression analyses were done with the Air Force data.

However, it was possible that the Air Force equations simply would not fit the simulation data. The subjects in the simulation were of higher ability, and although the learning situation in the simulation was similar to the Air Force setting, they were, of course, not exactly the same. Thus, equations were also constructed for the subjects in the simulation.

Four performance criteria were used as dependent variables in the regression equations. The first was total PT time. This is the total amount of time a subject took to complete the six target PTs, less the time taken to complete the appraisals. Since the time to take the appraisal is really testing time, it is not, strictly speaking, time on the PTs themselves. The second performance measure was the average of the scores on the six appraisals. It was based only on the score of the appraisals that were passed. That is, if a subject took the test and failed on his first attempt, but passed on his second, only his second score would enter the calculation of his mean appraisal score. The third measure was his score on the comprehensive test taken after the last PT was completed, and which covered the material on all six PTs. The final score was an overall performance measure composed of the three previous variables. This score was calculated for each subject by weighting the time to complete score twice as heavily as the sum of average appraisal score and the comprehensive test score. For this calculation, the following equation was used:

Composite Score =

$$2 \left[\frac{2000 - \# \text{ Minutes to Complete PTs}}{6 \text{ minutes}} \right] +$$

$$+ \frac{\frac{\text{Mean Appraisal Score}}{\sigma \text{ mean appraisal score}} + \frac{\text{Comprehensive Score}}{\sigma \text{ comprehensive score}}}{2}$$

Time to complete (reverse scored) was weighted more heavily since Air Force personnel felt speed was the performance variable of prime concern.

In order to deal with possible shrinkage, regression equations were developed using a double cross-validation procedure. The sample was randomly split into two equal groups, and build-up stepwise regressions were calculated for each sample, A and B. The weights developed in sample A were then applied to sample B, and the weights developed in B applied to A.

Results of regression analyses on the Air Force data are presented in Table 2. The development and cross-validation analyses are presented to each of the four performance variables. For each analysis, the specific ability tests entering the equation are presented in order of their entry into the equation. The resulting multiple R is also presented in order of their entry.

Inspection of the table indicated that prediction was generally quite good. The best overall index of predictability is the composite score, and for this measure, the multiple r for the total sample was .59. This compares very favorably with typical selection studies where predictive validities generally range in the .40s and .50s. The table also indicates that the equation are quite stable. Cross-validated Rs are quite close to the magnitude of the R's based on the development

Table 2
Ability Based Regression Equations Predicting Criteria
Using AF Sample

Performance Measure	Total Sample	Sample A	Sample B	B to A	A to B
	(N=187)	(N=90)	(N=97)		
Total PT Time	IQ .40**	IQ .45**	Para .33*		
	Study .43**	Study .48**	Mech .38*		
	Para .44**	Para .49**	Study .40*		
	Mech .44**	Spatial .49**	IQ .40*		
	Spatial .44**	Mech .49**	Spatial .40*	.46**	.38**
	(N=191)	(N=90)	(N=101)		
Average Appraisal Score	IQ .48**	IQ .50**	IQ .47**		
	Spatial .54**	Spatial .59**	Spatial .55**		
	Mech .54**	Mech .59**	Para .53**		
	Study .54**	Study .59**	Study .53**		
	Para .54**	Para .59**	Mech .53**	.57**	.51**

Comprehensive Test	(N=191)		(N=90)		(N=101)		
	IQ	.51**	IQ	.54**	IQ	.48**	
	Study	.52**	Spatial	.55**	Study	.50**	
	Mech	.53**	Para	.60**	Mech	.51**	
	Para	.54**	Mech	.61**	Spatial	.51**	
	Spatial	.54**			Para	.51**	.55** .44**
Composite	(N=187)		(N=90)		(N=97)		
	IQ	.52**	IQ	.57*	IQ	.51**	
	Study	.57**	Study	.60**	Mech	.54**	
	Mech	.58**	Para	.61**	Para	.56**	
	Para	.59**	Spatial	.61**	Study	.57**	
	Spatial	.59**	Mech	.61**	Spatial	.57**	.60** .56**

*p < .01

**p < .001

sample. The average shrinkage was only .03 correlation units.

Analogous data for the simulation sample are presented in Table 3. These analyses utilize the ability data from the simulation sample to predict performance in the simulation.

Table 3
Ability Based Regression Equation Predicting Criteria Using
Simulation Sample

Performance Measure	Total Sample	Development Sample	Cross Validated R's
	(N=57)	(N=39)	(N=18)
Total PT Time	IQ	Mech	
	Mech	IQ	
	Para	Study	
	Study	Para	
	Spatial		.36
	(N=57)	(N=39)	(N=18)
Average	Para	Para	
	Mech	Study	
	IQ	Mech	
	Study	IQ	
	Spatial		.68***
	(N=57)	(N=39)	(N=21)
Comprehensive Test	Para	Para	
	IQ	Spatial	

	Study .86 ***	Study .79 ***	
	Spatial .87 ***	IQ .79 ***	
	Mech .87 ***	Mech .79 ***	.90 ***
Composite	(N=57)	(N=39)	
	IQ .60 ***	IQ .59 ***	
	Mech .62 ***	Mech .64 ***	
	Study .63 ***	Study .64 ***	
	Spatial .63 ***	Para .65 ***	
	Para .63 ***	Spatial .65 ***	.63**

* p < .05
 ** p < .01
 *** p < .001

The cross validation procedure was somewhat different for these data. Due to the small sample size (N=57)² a double cross validation procedure was felt inappropriate. Thus, a traditional cross validation design was employed where two-thirds of the sample was randomly selected to constitute the development group, and the remaining third the hold-out group.

The data in this table indicate that the level of predictions in the simulation was almost identical to that in the AF sample. For example, the Air Force equations predicted total PT time .44 while the simulation data provided an R of .45. Analogous multiple correlations for the composite were .63 and .59. The simulation equations were also fairly stable. The only equation showing any real shrinkage was for total PT time.

However, even though the magnitude of the R's in the simulation data is comparable to those in the Air Force data, it should be noted that the order of the predictors entering the equations and the change in R at each step varies from the Air Force to the simulation data. For example, IQ was the best predictor in every case when the total Air Force sample was used, but only in two of the four cases for the simulation sample.

This indicates that the structure of the Air Force regression equation was different from that of the simulation equations. For this reason, it was felt necessary to more directly compare the two sets of equations. Table 4 presents data pertinent to one aspect of this comparison. The regression equations developed in the Air Force sample were applied to the ability data in the simulation sample and this predicted performance was correlated with actual performance. As the table indicates, use of the Air Force weights with the simulation data results in levels of predictability very close to those obtained when

²Since three of the 60 subjects did not complete PT6, their data could not be used. Thus, N=57.

the weights were applied to the samples from which they were generated. Thus, although the actual equations for the Air Force and simulation samples differed, the two sets of equations predict equally well in the simulation data.

A second method of comparison involves comparing the scores predicted by the Air Force equations with those predicted by the simulation equations. To do this, simulation subjects received predicted scores for a given performance measure: (1) based on the Air Force equation, and (2) based on the simulation equation. These two predicted scores were then correlated across the simulation subjects. The resulting correlations were:

Total PT time	.87
Mean Appraisal Score	.89
Comprehensive Score	.96
Composite	.98

Thus, the two equations produced almost the identical rank ordering of the simulation subjects.

Table 4
Comparison of Predictability of Criteria Using AF Weights Applied to AF and Simulation Samples

Performance Measure	AF Weights Applied to AF Sample	AF Weights Applied to Simulation Sample	Simulation Weights Applied to Simulation Sample
Total PT Time	.44*** (N=187)	.41** (N=57)	.45* (N=57)
Average Appraisal Score	.54*** (N=191)	.49*** (N=57)	.54** (N=57)
Score on Comprehensive Test	.54*** (N=191)	.83*** (N=57)	.87*** (N=57)
Composite Performance	.59*** (N=191)	.63*** (N=57)	.63*** (N=57)

* $p < .05$

** $p < .01$

*** $p < .001$

Generation of the Derived Effort Scores

The basic rationale in the derived effort measure is that if one partials ability out of performance, the resulting residual should be highly saturated with effort variance. However, the partialling procedure can be carried out in basically two ways, following either

additive or multiplicative assumptions. The first derived effort score assumed that effort and ability summed to equal performance. Consequently, the derived effort score was calculated by obtaining the predicted score for a subject on each of the four performance measures using the prediction equations previously generated. These, of course, were based solely on ability data, so the predicted score was essentially the level of performance predicted for a given subject on the basis of his level of ability. This score was subtracted from his actual performance score, and this residual constituted the derived effort measure. This procedure resulted in eight derived effort scores for each subject. One for each of the four performance variables using the Air Force weights, and one for each of the four performance variables using the simulation weights.

The second type of derived effort score was based on a model which states that performance is a function of ability multiplied by effort. In the previous additive model, $\text{Performance} = \text{Ability} + \text{Effort}$, thus $\text{Effort} = \text{Performance} - \text{Ability}$. Thus, subtracting predicted performance (ability) from performance is the derived effort score. However, in the multiplicative model, $\text{Performance} = \text{Ability} \times \text{Effort}$, thus, $\text{Effort} = \text{Performance}/\text{Ability}$. The derived effort score is thus actual performance divided by predicted performance (ability).

Derived effort scores based on this multiplicative model were calculated for each subject based on both the additive and multiplicative approaches. These were then compared to the hard criterion of effort.

Evaluating the Hard Criterion of Effort

Before turning to relationships between derived effort and the hard criterion of effort, it is appropriate to discuss data pertinent to the evaluation of the hard criterion of effort.

Recall that the effort data came from ratings of 8 mm photographs of the subjects. A frame was taken every six seconds. For rating purposes, every third frame was utilized. The frame was projected on a screen, and raters made a primary judgment as to whether the subject was working on the task or not for that frame. Subjects were rated for all the time that they were not actually taking an appraisal. Thus, if a subject was not in his seat in the picture, and he was not taking an appraisal, he was counted as not working for that frame. Also, since the material generally required eye-contact to work on, a subject was scored as not working if he was looking up from the work, or talking to a co-worker. Subjects had been told to work on the materials independently. This procedure resulted in 800-900 ratings per subject, per day.

The hard criterion of effort was then calculated for each subject as the number of frames he was working on the material, divided by the maximum number of frames he could have been working (i.e. appraisal time was removed). This "percentage of time on task" constituted the hard criterion of effort.

In evaluating the adequacy of this measure, several criteria were employed. The first was whether it produced variability. In fact it

did. The mean percent time on task was 79.7% with a standard deviation of 11.7. The range was from 48% to 96%. Clearly, variability was obtained.

The second criterion was the inter-rater reliability of the effort ratings. To assess this, the two raters independently rated 100 frames for 10 subjects from each of the three experimental conditions. The percent time on task was calculated for each of the 30 subjects, once for rater A and once for rater B. The difference between the percentages obtained by the two raters was then calculated and averaged across the 30 subjects. The mean difference in percentages was 3%. Thus, even with a fairly small sample of behavior, the ratings were highly reliable.

The third evaluation of the effort measure dealt with construct validity. If the effort measure is indeed a good one it should correlate significantly with actual performance, but since performance is determined by factors other than effort the correlation should be far from perfect. The effort measure correlated $-.44$ with a total PT time, $.14$ with average appraisal score, $.39$ with comprehensive test score, and $.49$ with the composite. (Note that a negative correlation with total PT time was expected since the greater the effort, the less time it should take to finish PTs). The magnitudes of the correlations of effort with the primary performance variables, total PT time and the composite, are in the expected range and thus add support to the validity of the effort measure.

Additional evidence of validity could be assessed by comparing effort scores across the 6 PTs, and with the performance measures. We would predict that: (1) effort scores should correlate highly across PTs; (2) correlations between PTs should be higher than correlations of effort with performance. The average correlation between effort scores across different PTs was $.58$. Thus, effort scores are fairly highly correlated, and are correlated higher with each other than with performance measures.

Self ratings of effort were also obtained from the subjects when they had finished the six AF PTs. Two items were utilized. The first asked "on this job I am working: ... As hard as I possibly can... About average ... I am taking it easy." A nine-point Likert response format was utilized with verbal anchors at every other step. (See Appendix D for the actual items). The second question asked "In terms of the total amount of effort I could put in on this job, I am putting in about: ...10% effort...50% effort...90% effort." As before, a nine-point scale was used. The sum of the responses to these two items constituted the self rating of effort index.

The central issue here is how well the hard criterion of effort was related to the self ratings of effort. The correlation between the hard criterion and the self rating was $.21$. While this is statistically significant ($p < .05$), it is quite small.

In theory, this could cast doubt on the validity of the hard criterion of effort. However, a more parsimonious explanation is that the self ratings were not particularly valid. The principal reason for this conclusion is that the self ratings actually correlated more

highly with total PT time ($r = .27$) than with effort. Apparently, the subjects were not considering their actual level of energy expenditure when responding to the items. Furthermore, self ratings of effort have been shown in a similar context (Pritchard, Von Bergen, and DeLeo, 1974) to have very low convergent validity.

Predictive Validity of the Derived Effort Measure

The primary method of evaluating the utility of the derived effort measure is to correlate the derived effort measure with the hard criterion of effort. Table 5 presents these correlations for the derived effort scores based on each of the four performance dimensions, for both the additive and multiplicative models, and for Air Force and simulation generated weights.

The overall conclusion from these correlations is that the derived effort measure does not predict actual effort particularly well. For example, correlations based on the composite ranged from $-.02$ to $.32$. The table also indicates that the scores based on the multiplicative model did no better than those based on the additive model. In three cases the multiplicative was better, in two cases the additive was better, and in three cases they were equal. Furthermore, none of the differences was of appreciable magnitude.

One clear finding is that when the derived effort scores are based on the regression equations calculated from the simulation data, they predict effort better than when the derived effort scores come from

Table 5
Correlations Between Derived Effort and Actual Effort (N=57)

Derived Effort Measure	Air Force Weights		Simulation Weights	
	Additive Model	Multiplicative Model	Additive Model	Multiplicative Model
Derived Effort - Total PT time	-.09	-.18	-.29*	-.30*
Derived Effort - Average Appraisal Score	-.27*	-.27*	-.03	-.04
Derived Effort - Comprehensive Score	.23*	.24*	.13	.13
Derived Effort - Composite	.16	-.02	.32**	.32**

* $p < .05$

** $p < .01$

the independent Air Force sample. For example, for the additive model the composite derived effort scores calculated from Air Force weights

correlated .16 with effort, but this value increased to .32 when the simulation-based derived effort scores were used. This occurred in spite of the fact that those two derived effort scores displayed a correlation between themselves of .98. One explanation for this pattern of results is based on the fact that the equations in fact predict lower performance for a given subject than do the simulation equations. This mean difference would not of course, affect the correlation between the two predicted scores. Thus, the two equations rank order the subjects almost exactly the same, but the Air Force prediction is lower. This, of course, would result in different derived effort scores since this predicted score is subtracted from actual performance. If, in fact, the regression line between predicted and actual performance in the Air Force data was parallel to that in the simulation data, the two sets of derived effort scores would differ only by a constant, and thus be equally correlated with actual effort. However, the two regression lines are not parallel. One explanation for this could be that since there were more high ability subjects in the simulation, this may have influenced the regression line. This would imply that the relationship between ability and performance is non-linear. This will be discussed in more detail later.

Another way of assessing the utility of the derived effort measure is to deal with it at a more gross level than the accuracy of individual prediction. Recall that in order to produce variability in effort, three experimental conditions were employed, Hourly, Fixed Ratio, (FR) and Variable Ratio-Variable Amount (VRVA) pay systems. These three pay conditions did, in fact, produce variability in performance. The issue is how well the derived effort measures discriminated the three conditions. If the derived effort is useful, the three conditions should show even a greater difference in effort than they do in performance. This is the case since performance contains variance due to ability, but mean ability was constant across the three conditions.

Consequently, one way ANOVAS were calculated using the three conditions as levels of the factor and : (a) performance as the dependent variable, and (2) derived effort as the dependent variable. Since the pay system only dealt with number of PTs passed rather than appraisal scores or score on the comprehensive test, only the total PT time variable was appropriate for these analyses.

The resulting F-ratio for actual time to complete the PTs was 7.29 ($p = .002$). The F for the derived effort analysis using Air Force weights was 11.62 ($p < .000$), and using simulation weights was 11.89 ($p < .000$). Since the order of the actual means was in the same direction for all three analyses, the larger F-ratios for the derived effort analyses indicate that using derived effort does in fact result in less error.

In examining the scatterplots relating derived effort to actual effort it was observed that there were a number of subjects which exhibited a specific pattern of scores. They were subjects of relatively low ability who finished the material quite quickly, but who spent a relatively small percentage of time actually working on the task. One explanation for this pattern of scores was that these subjects were obtaining the answers to the appraisals from other people.

During the conduct of the simulation it became apparent that this was actually occurring to some extent.

To deal with this potential problem, a "cheating index" was formed. It was based on the assumption that someone who finished the last of the six Air Force PTs substantially faster than they finished the first PT, and who did poorly on the comprehensive test was probably getting the answers to the appraisals from someone else. The logic here was that it was unlikely that a subject would get the answers from someone else on the first PT, before they were accustomed to the situation. Thus, their time to complete the first PT could be viewed as their actual level of performance. If, by the sixth PT they were getting the answers, they should be able to complete it much faster. If, however, they did not actually know the material, this should show up as a low score on the comprehensive test taken after the last Air Force PT. Since no specific score was required to "pass" the comprehensive test there would be no pressure to pass answers for this. Furthermore, these tests were not scored during the simulation, so a subject would have no knowledge of what the correct answers were, and thus passing them to someone else was not possible.

Thus, to construct this cheating index, data from the relative time taken to complete PT 1 and PT 6, and the comprehensive scores were examined. The first step was to calculate the percent change in time to complete PT 1 and PT 6. Examination of this distribution indicated that there seemed to be a break in the distribution at about the 30% point. Thus, subjects who were more than 30% faster on PT 6 than on PT 1 were considered potential cheaters. The comprehensive scores for these subjects were then examined, and any of these subjects who received a score of less than 75% on the comprehensive was considered a cheater. This procedure resulted in the elimination of 10 subjects. A second criterion was also employed. If a subject did very poorly on the comprehensive test (less than 60%) and their percent increase in speed from PT 1 and PT 6 was positive, they were considered a cheater. This criterion eliminated one additional subject.

Thus, the procedure resulted in the elimination of 11 subjects. It is likely that some of these subjects were not, in fact, receiving answers but it was felt better to eliminate a non-cheater than retain a true cheater. Some evidence for the validity of the cheating index is available in that there were four subjects who were known by the instructors to have been cheating. All four were included in the 11 subjects eliminated by the cheating index. Furthermore, all but one of the 11 eliminated were in the conditions where pay was dependent on number of PTs passed. It was for these conditions where passing a large number of PTs was financially worthwhile.

These subjects were removed, and the principal analyses repeated on the remaining 44 subjects. Regressions were calculated, derived effort scores were computed, and correlations between derived effort and actual effort recomputed. Table 6 presents the results of these analyses for the additive model derived effort scores, for both Air

Table 6
Correlations Between Derived Effort and Actual Effort with Cheaters
Removed and Included

Derived Effort Measure	Correlation Between Derived Effort and Actual Effort Total Sample (N=58)		Cheaters Excluded (N=44)	
	AF Weights	Simulation Weights	AF Weights	Simulation Weights
Derived Effort - Total PT Time	-.09	-.29*	-.40**	-.39**
Derived Effort - Average Appraisal Score	-.27*	-.03	-.26*	-.04
Derived Effort - Comprehensive Score	.23*	.13	-.08	-.02
Derived Effort - Composite	.16	.32**	.37**	.35**

* $p < .05$

** $p < .01$

Force and simulation weights. The analogous correlations for the total sample (cheaters included) is also repeated for comparison purposes. The table indicates that when the cheaters are excluded, the predictive validity of the composite derived effort increases somewhat -- from .32 to .37 using simulation weights. More importantly, the data suggest that when cheaters are removed, the composite derived effort based on Air Force weights predicts equally well as does that based on simulation weights. Thus, the difference in the predictability of effort found in the original analysis is probably due to the presence of cheaters in the simulation.

The ANOVAs comparing actual performance (Total PT time) and the derived effort scores across the three conditions were also repeated with cheaters excluded. The F-ratio for actual PT time was 3.86 ($p = .029$), for derived effort (additive model) using AF weights it was 5.69 ($p = .007$), and when simulation weights were used in was 7.49 ($p = .002$). As with the comparable analysis with the total sample, using derived effort results in more precision.

Although these results were more encouraging than the original analyses, the level of predictability of effort was still low. However, upon examining the scatterplots of derived effort with actual effort once the cheaters were removed, it was observed that subjects with low ability tended to be outliers from the main clustering of points. They generally tended to have derived effort scores higher than their actual effort. This suggested that actual level of ability might moderate

the relationship between derived effort and actual effort.

Consequently, the sample was split into high and low ability groups on the basis of the simulation equation scores. Subject's whose predicted performance (ability) was above the median were considered high ability, those below the median were considered low ability.

Regression equations were then developed for the two groups separately. Due to the small sample size ($N = 22$ per group) only two predictors were used in these equations. These were the two best predictors from the total sample. Derived effort scores were then calculated for the subjects in each group based on the equations for that group. Derived effort scores were calculated for only the two major performance variables, total PT time and the composite.

Results of these analyses are presented in Table 7. As the size of the correlations indicates, predictability is substantially increased when the sample is broken down by ability. This implies that somehow ability and effort are combining differently for high and low ability subjects.

Table 7
Correlations Between Derived and Actual Effort,
by High and Low Ability Subgroups

Derived Effort Measures	Correlation Between Derived and Actual Effort	
	Low Ability (N=22)	High Ability (N=22)
Derived Effort - Total PT Time	-.55**	-.43*
Derived Effort - Composite	.54**	.44*

* $p < .05$

** $p < .01$

Summary of the Results

1. The simulation subjects were roughly comparable to the Air Force subjects. The major difference was that there were more high ability subjects in the simulation sample, but both groups showed good variability.
2. The ability test battery predicted performance quite well. The composite performance index was predicted .59 in the Air Force sample, .63 in the simulation sample, and .63 when the Air Force weights were applied to the simulation data. The degree of index shrinkage was quite low.
3. The structure of the Air Force equations differed from the simulation equations, but they predicted simulation performance

equally well, and the predicted scores from the two equations were highly correlated.

4. The hard criterion of effort displayed wide variability, excellent reliability and good construct validity.
5. When the entire sample was used, derived effort did not predict actual effort particularly well.
6. Derived effort scores based on the multiplicative model did not predict effort any better than derived effort scores based on the additive model.
7. Derived effort scores discriminated the three experimental conditions better than actual performance.
8. When cheaters were removed from the sample, prediction of effort increased somewhat.
9. When separate analyses were conducted on high and low ability subsamples, prediction of effort was fairly good. The correlation for high ability subjects was .44, and .54 for low ability subjects.

Discussion and Conclusions

Given the results presented above, the major issue now becomes how well did the derived effort technique work, and under what circumstances could it be useful. It is to these issues we now turn.)

It is clear from the data that the derived effort measure produced statistically significant correlations with actual effort in almost every case. However, the presence of statistical significance is not enough to justify utilization of the technique. The actual magnitude of the relationship must be considered.

In considering this issue, we shall assume that the best estimate of the correlation between predicted and actual effort comes from the sample with the "cheaters" removed. There is good evidence that at least some of the subjects were in fact getting the answers elsewhere and, as such, they add artificial error variance. With this in mind, our best estimate of the relationship between derived effort and actual effort is obtained by correlations between the composite derived effort measure and actual effort. These relationships were in the mid to upper 30's.

A relationship of this magnitude is not large enough to use for individuals. That is, the derived effort index contains too much error to be used for making decisions about a given individual. However, it would be useful for group data. That is, if one group had a substantially higher derived effort score than another, it would be fairly safe to conclude that the group with the higher mean derived effort score was exerting higher effort. We are essentially arguing

that with a group, the error associated with each individual's derived effort score should be randomly distributed across the members of the group. As a consequence, the mean derived effort score across the group should reflect its average effort.

The data are more encouraging when the analyses broken down by high and low ability are considered. In these analyses, correlations between predicted and actual effort ranged from the mid .40's to the mid .50's. Correlations of this size are approaching a level of magnitude that might be useful for individual predictions. They still contain a large amount of error, and using the derived effort measure with such relationships would have to be done with caution. However, relationships of this magnitude would be extremely useful for groups.

The real issue here is whether the larger relationships found when the sample was broken down into high and low ability would, in fact, replicate in another sample. On a post hoc basis it is not unreasonable to argue that they would. It seems quite possible that ability and effort do in fact combine somewhat differently for low and high ability trainees. It remains to be seen, however, whether the finding would replicate.

In the introduction to this report, four major advantages were claimed for the derived effort index. It was argued that the technique: (1) would be an individualized procedure; (2) when used in an incentive system rewards could be given on the basis of effort rather than performance; (3) if used in an incentive system, trainees would have an equal chance to get incentives regardless of their ability; and (4) would be a useful measure of effort with which to compare different groups. We shall now discuss each of these potential advantages in the context of the results obtained from this research.

(1) Individualized. The derived effort technique is indeed individualized. It utilizes data from the individual. His ability is considered, and the predicted score generated from his ability data is compared to his actual performance. However, whether a technique is individualized per se is really a matter of instructional philosophy. The technique must also have other utility if one is to argue for its use.

(2) Reward can be based on effort. The data suggest that if the derived effort measure was used as an index of individual effort, and incentives were awarded on the basis of the derived effort scores for individuals, there is too much error in the system to enable one to say that rewards were in fact given on the basis of effort. Thus, the data indicate that when used for individuals the technique will not give rewards based on effort. However, if groups are given incentives on the basis of mean derived effort scores, one could argue that rewards were based on effort. This argument rests on the reasonable assumption that errors of predicting actual effort from derived effort would be random, and thus mean differences between groups reflect actual differences in effort. This is especially true if scores are based on high and low ability groups separately.

(3) Trainees have an equal chance to get incentives. Regardless of whether the derived effort score is a good measure of effort, it still has the definite advantage of tending to equalize the chances for earning incentives. Since high ability trainees have higher predicted scores than low ability trainees, the higher ability people must perform at a higher level to obtain incentives. Thus, lower ability students should see that obtaining incentives is more within their power than in the situation where actual raw performance is rewarded. Thus, the derived effort score is useful for those groups.

(4) Useful for comparing different groups. Clearly, the derived effort technique would be useful for comparing effort in different groups. For example, if a new instructional technique was introduced, and one wished to compare the level of motivation produced by the two techniques, the derived effort index would be an excellent way to do this. This is especially true when the actual level of ability of the samples exposed to the two techniques differed.

More importantly, the derived effort index gives one a common metric with which to compare different courses. In many cases it would be possible to directly compare different courses whose content, examinations, and formats were quite different. We shall consider this in more detail below.

Suggestions for Implementation

Thus far we have been considering ways in which the derived effort index could be used in only a very general way. The conclusion is that it is a very useful index for group effort, but probably not appropriate for individual assessment. Within this restriction, however, it has a number of very practical uses. The research was originally conducted with a view toward the Advanced Instructional System (AIS), and these applications to be discussed apply directly to the AIS. However, many of them could be used in any training course.

(1) Within a given course. A very useful application of the derived effort index would be to compare different parts of a given course. The issue is whether different parts (e.g., blocks) of course result in greater levels of motivation than others. If this could be determined, it would be of great aid in redesigning courses. The procedure would basically involve generating the predicted scores for total course performance and then breaking it down by block. Let us use time to complete the course as the criterion of interest for an example, although exam scores or any other criterion could be used. One would first predict speed of completion of the entire course from the ability data. However, since the blocks vary in amount of material, the prediction of speed of completion must consider this. Suppose, for example, there were three blocks. Based on available data trainees average 20 days on the first block, 30 days on the second, and 50 days on the third. Thus, average time to complete is 100 days. The first block represents 20% of total time; the second, 30%; and the third 50%. Thus, the predicted time to complete the course for a given subject

would be proportioned according to these percentages to obtain a prediction of his time to complete each block. If a given subject received a predicted score of 110 days, we would predict he would complete block 1 in 22 days ($20\% \times 110$), block 2 in 33 days, and block 3 in 55 days. His actual time to complete each block would then be subtracted from his predicted time.

Once this is done for all students, the mean derived effort by block could be calculated. If the three blocks result in equal motivation, the derived effort index should be near zero for all blocks. The three derived effort means are substantially different, it would indicate that differences in motivation were present.

This procedure could be refined by doing separate analyses for high and low ability students. The results would indicate whether high and low ability students exhibited the same pattern of motivation across blocks.

(2) Changes in a course. Another application of the derived effort index is where a change is made in the structure or procedures in a given course. If students in the original format of the course have ability levels equal to those in the course after it has changed, simple performance measures could be used. However, as the ANOVAs in our data have indicated, using the derived effort index gives a more precise test of effects since ability variation within conditions is controlled. The real advantage to the derived effort index comes about when the actual ability of the two samples is not equal. Then, the derived effort index would be very useful.

(3) Comparisons between different courses. As we have discussed above, the derived effort index could be very useful in comparisons between different courses. One procedure would be to simply calculate derived effort for the two courses and compare the two means. However, if the regression equations are based on the same sample upon which the derived effort scores are calculated, both means should be zero. However, when one or the other courses is changed in some way, the derived effort index should be able to detect changes relative to the other course. One could assess, for example, whether a technique used in one course had an equal impact on another course.

Another procedure would be to go back to data obtained from students who had completed the course at a given time and develop the regression equations on that sample. If these equations were applied to another (e.g., more recent) sample, differences in mean derived effort would be more interpretable.

(4) Feedback Another application of the derived effort index would be for feedback purposes. Instructors could be given the mean derived effort score for each section they taught and the change in this value from class to class could be useful information. Students could also be given such feedback on a group basis, and it could provide them information about their motivation.

(5) Goal setting. One could also use the derived effort index on an individual basis in the context of goal setting. As we have argued above, the derived effort index for a single individual is not an accurate index of his own effort. However, if the individual's

predicted score was presented to him as a basis for goal setting, it would at least have meaning in the sense that it represents the average performance for people of comparable ability. An easy goal could be to meet the predicted score, a hard goal could be to perform at a level one standard deviation above the predicted goal. In fact, one could readily list a number of performance goals and indicate the objective probability (based on the development sample) of obtaining that goal.

Overall, then, the results of the present research indicate that while the derived effort index should probably not be used as a measure of effort for a single individual, the index can be considered a measure of group effort. As such it has a number of very useful applications in both the AIS and other training contexts.

References

- Appelzweig, M.H., Moeller, G., & Burdick, H. Multi-motive prediction of academic success. Psychology Reports, 1956, 2, 489-496.
- Atkinson, J.W. An introduction to motivation. Princeton, N.J.: Van Nostrand, 1964.
- Barnes, R.M. Motion and time study. New York: Wiley and Sons, 1940.
- Bitterman, M.E. Heart rate and frequency of blinking as indices of visual efficiency. Journal of Experimental Psychology, 1945, 35, 279-292.
- Braunstein, D.H., Bruanstein, H.M., & Blumfield, W.J. Performance in training and an achievement effort rating. Psychology Reports, 1965, 16, 1077-1080.
- Campbell, J.C., Dunnette, M.D., Lawler, E.E., & Weick, K.E. Managerial Behavior, performance, and effectiveness, McGraw-Hill, 1970.
- Davis, R.C. Patterns of muscular activity during mental work, and their constancy. Journal of Experimental Psychology, 1939, 24, 451-465.
- Farquhar, W. H. A comprehensive study of the motivational factors underlying achievement of eleventh grade high school students. East Lansing: Michigan State University, 1963.
- Ford, F.A. The ratio of achievement to ability as found among fifth-grade pupils. Contribution to Education, No. 94, George Peabody College for Teachers, Nashville, Tenn. 1931.
- Furst, E.J. Validity of some objective scales of motivation for predicting academic achievement. Education and Psychological Measurement, 1966, 26(4), 927-933.
- Hackman, R., & Porter, L.W. Expectancy theory predictions of work effectiveness. Organizational Behavior and Human Performance, 1968, 3, 417-426.
- Haven, S.E. The relative effort of children of native versus foreign born parents. Journal of Educational Psychology, 1931, 22, 523-535.
- Lawler, E.E. Pay and managerial effectiveness: A Psychological view. New York: McGraw-Hill, 1971.
- Lawler, E.E., & Porter, L.W. Antecedent attitudes of effective managerial performance. Organizational Behavior and Human Performance, 1967, 2, 122-142.

- Luekel, F. Introduction to psychological psychology. Saint Louis: C.V. Mosby, 1968.
- Lovekin, O.S. The quantitative measurement of human efficiency under factor conditions. Journal of Industrial Hygiene and Toxicology, 1930, 12, 99-120.
- Luchins, A.J., & Luchins, E.H. The Einstellung phenomenon and effortfulness of task. Journal of General Psychology, 1954, 50, 15-27.
- Mayo, R.J. The conceptual representation and measurement of human effort. Unpublished paper, Purdue University, 1974.
- Mayo, G.D., & Manning, W.H. Motivation measurement. Education and Psychological Measurement, 1961, 21 (1), 73-83.
- McCall, W.A. How to Experiment in Education. New York: Macmillan, 1930.
- McClelland, D.C. Studies in Motivation. New York: Appleton-Century-Croft, Inc., 1955.
- Mitchell, V.F. The relationship of effort, abilities, and role perceptions to managerial performance. Unpublished doctoral dissertation, University of California, Berkeley, 1966.
- New York Institute of Technology, A programmed text in basis electricity. New York: McGraw-Hill, 1963.
- New York Institute of Technology, A programmed text in basic electronics. New York: McGraw-Hill, 1964.
- Pinter, R. Twenty first yearbook, 1930.
- Porter, L.W. & Lawler, E.E. Managerial attitudes and performance. Homewood, Ill.: Irwin Dorsey, 1968.
- Presgrave, R. The dynamics of time study. Toronto: University of Toronto Press. 1954.
- Pritchard, R.D., Leonard, D.W., Von Bergen, C.W. Jr., & Kirk, R.J. The effects of varying schedules of incentive delivery on technical training AFHRL-TR-74-32, Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado, (1974). AD- A001117
- Pritchard, R.D., Von Bergen, C.W. Jr. & DeLeo, P.J. Incentive motivation techniques evaluation in Air Force Technical Training AFHRL-TR-74-24, Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado, (1974). AD-A005302.

- Ramby, S.M. The evaluation of an indirect technique for assessing effort. Unpublised M.S. thesis, Purdue University, 1973.
- Ryan, T.A. Work and effort. New York: Ronald Press, 1974.
- Schneider, B., & Olsen, L.K. Effort as a correlate of organizational reward system and individual values. Personnel Psychology, 1970, 23, 313-326.
- Scott, S.G. Auditory flutter fussion as a measure of effort. Dissertation Abstracts International, 1960, 21, 1926.
- Solomon, R.L. Time and effort factors in the avoidance of repetition of responses. American Psychologist, 1946, 1, 291-292.
- Strauss, G., & Sayles, L.R. Personnel: The Human Problems of Management. New York: Prentice-Hall, 1960.
- Thorndike, E.L. Educational psychology; Volume I, the original nature of man. New York: Columbia University Press, 1913.
- Tsao, F. Is AQ or F score the last word in determining individual effort? Journal of Educational Psychology, 1943, 24 (9), 513-525.
- Vroom, V. Work and motivation, New York: Wiley, 1964.
- Yacorzynski, G.K. Degrees of effort II: quality of work and time of completion of performance tests. Journal of Experimental Psychology 1942, 30, 342-344.

APPENDIX A

Advertisement Used to Recruit Subjects

HIGH SCHOOL STUDENTS ONE WEEK JOB

We are looking for 20 people between 17 - 19 years of age to work for one week on a job evaluating written training materials. No special skills are required. The pay will be approximately \$2.00 per hour, depending on what you do. The work day will be from 8:30 A.M. to 1:30 P.M., Monday through Friday, June 17 - 21.

If you would like a week's work, report at 8:30 A.M., Monday, June 17 at the conference room in the Holiday Inn, U.S. 24 East, Logansport, Indiana.

APPENDIX B

Pre-employment Electricity Test

1. What type of aircraft is a KC-135A?

2. What is galvanic corrosion?

Define the following terms and symbols:

3. ampere

4. static discharger

5. magnetic permeability

6. E

7. 

8. buttock lines

9. multimeter

APPENDIX C

Sample Appraisal

APPRAISAL*

COURSE: Aircraft Electrical Repairman

SUBJECT: DC Generation and Basic Circuit Symbols and Terms

INSTRUCTIONS: Follow the directions given in each section.

Section I

OBJECTIVE: Given the names of electrical components, identify each component that belongs to one of the following categories: a. source of EMF, b. protective devices, c. control devices, d. load devices, and e. conductors. A minimum of 80% accuracy is required.

Match the terms on the right to the components on the left. Place the letter of the term in the blank provided by the component. The terms may be used more than once.

COMPONENTS

TERMS

- | | |
|---------------------------|----------------------|
| ___ 1. Fuse | a. Conductor |
| ___ 2. Motor | b. Load Unit |
| ___ 3. Lamp | c. Source of EMF |
| ___ 4. Circuit Breaker | d. Control Device |
| ___ 5. Generator | e. Protective Device |
| ___ 6. Thermocouple | |
| ___ 7. Resistor | |
| ___ 8. Aircraft Structure | |
| ___ 9. Battery | |
| ___ 10. Switch | |

Section II

OBJECTIVE: Given a list of electrical symbols and a list of units and terms, match the symbols with their respective unit or term. A minimum of 80% accuracy is required.








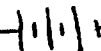

*The above appraisal has been discontinued.

Match the terms with the symbols below. Place your answer in the blank provided by the term.

TERMS

- | | | |
|------------------|--------------------------|-----------------|
| ___ 1. Fuse | ___ 6. Circuit Breaker | ___ 11. Voltage |
| ___ 2. Lamp | ___ 7. Fixed Resistor | ___ 12. Current |
| ___ 3. Battery | ___ 8. Variable Resistor | ___ 13. Amperes |
| ___ 4. Ammeter | ___ 9. Thermocouple | ___ 14. Ohms |
| ___ 5. Generator | ___ 10. Resistance | ___ 15. Volts |

SYMBOLS

- | | | | | |
|--|--|--|------|------|
| a.  | d.  | g.  | j. R | n. I |
| b.  | e.  | h.  | k. a | o. |
| c.  | f.  | i.  | l. E | |
| | | | m. V | |

Section III

OBJECTIVE: Given a list of definitions and a list of DC generation terms match the definition with their proper terms. A minimum of 80% accuracy is required.

Match the definitions on the right to the terms on the left, and record your answers in the blanks provided by the terms

TERMS

DEFINITIONS

- | | |
|-------------------------------|--|
| ___ 1. Generator | a. Voltage produced when heat is applied to two dissimilar metals that are jointed at one end. |
| ___ 2. Battery | b. Device using mechanical energy to produce an EMF |
| ___ 3. Thermocouple | c. Device using heat to produce an EMF |
| ___ 4. Mechanical Method | d. Device using the chemical method to produce an EMF |
| ___ 5. Heat or Thermal Method | e. Voltage produced by relative angular motion between conductors and a magnetic field |

APPENDIX D

Effort Questionnaire

PART II

Circle the number that best describes your feelings. Circle any number. If you feel you are, for example, between the statement in number 7 and the statement in number 5, circle number 6.

1. On this job I am working

9. As hard as I possibly can

8.

7. Fairly hard, but not killing myself

6.

5. About average

4.

3. Not very hard

2.

1. I am taking it easy

2. In terms of the total amount of effort I could put in on this job, I am putting in about:

1. 10%

2. 20%

3. 30%

4. 40%

5. 50%

6. 60%

7. 70%

8. 80%

9. 90%